The Objection to Claims Under 35 USC 102

The invention by Iizuka requires a system operator to enter parser metadata for each html page to indicate how to extract data from the page. The purpose of the present invention is to eliminate that requirement.

The Objection to Claims Under 35 USC 103

The patent by Zimmerman et. al. teaches a system that searches files for words in order to index or classify files. The system does not extract records from files.

General

The present invention extracts records from sources. The following cited references do not relate to methods to extract records from sources: Bull et. al., Zimmerman et. al., Smith et. al., Kirk et. al., Iijima, Yokoyama et. al.

The present invention automatically deduces how to parse sources. The following cited references do not relate to methods to automatically deduce how to parse sources: Hammer et. al., Iizuka et. al., Ambite et. al., Temkin et. al., Kiuchi at. Al., Stern, Rajan et. al.

The present invention is related to the methods described in the papers "Wrapper Generation for Semi-Structured Internet Sources" and "Semi-automatic Wrapper Generation for Internet Information Sources," both by Ashish and Knoblock. (Thanks to the examiner for bringing these papers to the attention of the applicant.) Both the present invention and those methods have the aim of extracting data from semi-structured sources. A fundamental difference is that the present invention uses recognition of potential data values as well as labels, while the methods by Ashish and Knoblock use recognition of labels alone. Recognizing potential data values enables the following advantages:

- The invention can be applied to data sources that have multiple records per page, for example, data sources that have tabular data with one row per record and each field labeled only in the header row of the table rather than within each record row. This is a common format for records in html sources.
- The invention can be applied to data sources in which the text corresponding to each record is free text, for example, where the text corresponding to multiple fields of each record is in the same element of an html document. This format is common when users enter free text information about items for sale, especially when the items are used or one-of-a-kind. Different users enter data for different record fields in different orders, and there is no clear separation between text corresponding to different data fields.

The present invention is also related to a method by Kushmerick, described in "Wrapper induction: Efficiency and expressiveness (Extended abstract)." (The applicant found this work through references in a paper by Ashish and Knoblock. Once again thanks to the examiner for locating the work by Ashish and Knoblock.) Both the present invention and the work by Kushmerick have the aim of extracting data from semi-structured sources. The method by Kushmerick deduces a parser that extracts records. The deduction is based on having a few examples of sources with the same structure as the source to be parsed, and the user has to supply the records that would be the outputs of a successful parser for the examples of sources. Hence, a user must perform the extraction by hand on examples in order for a computer to generate the parser. The present invention does not have this requirement. Also, the limitations of the parsers that can be generated by the method by Kushmerick cause the method to fail for the following cases:

- sources in which the text corresponding to each record is free text, for example, where the text corresponding to multiple fields of each record is in the same element of an html document, and
- sources in which the text corresponding to different data fields is in different orders within the text corresponding to different records.

The present invention can operate in these cases because it uses recognition of potential data values.

Note on Claims

The new claims correspond to combinations of the old claims as follows. The new Claim 1 contains elements related to the old Claims 1, 4, 5, 8, 9, and 10. The new Claim 2 contains elements related to the old Claim 6. The new Claim 4 contains elements related to the old Claims 15, 18, 19, 22, 23, and 24.

Conditional Request for Constructive Assistance

If the application is not believed to be in full condition for allowance, we request further constructive assistance and suggestions from the examiner.

Eric T. Bax

PO Box 60543
Pasadena, CA 91116-6543

626-827-0446

Charless C. Fowlkes

31 Gardner Park Drive
Bozeman, Montana 59715

Louis Cisnero Jr.

1208 Commerce Street
Jourdanton, Texas 78026